Am. J. Hum. Genet. 66:1328–1340, 2000

Detection of Disease Genes by Use of Family Data. I. Likelihood-Based Theory

Alice S. Whittemore and I-Ping Tu

Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA

We present a class of likelihood-based score statistics that accommodate genotypes of both unrelated individuals and families, thereby combining the advantages of case-control and family-based designs. The likelihood extends the one proposed by Schaid and colleagues (Schaid and Sommer 1993, 1994; Schaid 1996; Schaid and Li 1997) to arbitrary family structures with arbitrary patterns of missing data and to dense sets of multiple markers. The score statistic comprises two component test statistics. The first component statistic, the nonfounder statistic, evaluates disequilibrium in the transmission of marker alleles from parents to offspring. This statistic, when applied to nuclear families, generalizes the transmission/disequilibrium test to arbitrary numbers of affected and unaffected siblings, with or without typed parents. The second component statistic, the founder statistic, compares observed or inferred marker genotypes in the family founders with those of controls or those of some reference population. The founder statistic generalizes the statistics commonly used for case-control data. The strengths of the approach include both the ability to assess, by comparison of nonfounder and founder statistics, the potential bias resulting from population stratification and the ability to accommodate arbitrary family structures, thus eliminating the need for many different ad hoc tests. A limitation of the approach is the potential power loss and/or bias resulting from inappropriate assumptions on the distribution of founder genotypes. The systematic likelihood-based framework provided here should be useful in the evaluation of both the relative merits of case-control and various family-based designs and the relative merits of different tests applied to the same design. It should also be useful for genotype-disease association studies done with the use of a dense set of multiple markers.

Introduction

In some diseases with complex genetic etiologies, conflicting results have emerged from case-control studies of association, compared with linkage analyses based on allele-sharing within families. Specifically, although the case-control studies have shown strong associations, the linkage tests have proved negative (Parsian et al. 1991). To explain this phenomenon, Risch and Merikangas (1996) have suggested that allele-sharing linkage tests can have poor power compared with tests for association and that a genomewide search for associations may be more sensitive than genome scanning for determination of linkage.

However, case-control studies may give biased measures of association as a result of unrecognized ethnic admixture of the population (known as the "population stratification" problem). This possibility has prompted interest in the use of family-based designs. Comparison of genotypes of affected individuals with those of their unaffected siblings or with Mendelian expectation based on the genotypes of their parents allows such designs to avoid this problem. However, family-based designs can be less powerful than case-control designs (Witte et al. 1999), and their advantage is unclear in light of uncertainty about the need to control for population stratification (Rothman et al. 1999).

In the present study, we derive a class of likelihood-based test statistics that are applicable to cases, controls, and arbitrary families with arbitrary patterns of missing data and that combine the advantages of family-based and case-control designs. The tests are based on the score statistics derived from a specific likelihood for the data. The likelihood function extends, to arbitrary families and to multiple markers, the likelihood proposed by Schaid and co-authors (Schaid and Sommer 1993, 1994; Schaid 1996; Schaid and Li 1997) for nuclear families with only affected offspring.

The score statistic comprises two component test statistics. The first statistic, the *nonfounder statistic* (NFS), evaluates transmission disequilibrium from parents to offspring. This statistic generalizes the transmission/disequilibrium test (TDT) (Ott 1989; Terwilliger and Ott 1992; Knapp et al. 1993; Spielman et al. 1993; Ewens and Spielman 1995; Spielman and Ewens 1996) and the

Received June 1, 1999; accepted for publication January 18, 2000; electronically published March 29, 2000.

Address for correspondence and reprints: Dr. Alice S. Whittemore, Department of Health Research and Policy, Stanford University School of Medicine, Redwood Building, Room T204, Stanford, CA 94305-5405. E-mail: alicesw@leland.stanford.edu

0002-9297/2000/6604-0016$02.00

score statistics proposed by Schaid and Sommer (1994), to include multiallelic markers, markers distinct from the trait locus, and multiple markers, with the use of families with both affected and unaffected offspring and families with missing parental genotypes. The second component statistic, the *founder statistic* (FS), compares marker genotypes in the family founders with those expected under the null hypothesis. This statistic generalizes the statistics that are commonly used for case-control data (Barcellos et al. 1997; Risch and Teng 1998).

We illustrate the tests with some simple examples. These focus on genotypes at a single diallelic marker that is in partial or complete disequilibrium with the etiologically relevant disease locus. In the present study, which is published with a companion article (Tu et al. 2000) in this issue of the *Journal,* we apply the statistics to the special case of unrelated individuals, whereas, in the companion article (Tu et al. 2000), we treat nuclear families. Basing our tests on a likelihood function requires that we specify a penetrance model for the relationship between the disease and the genotypes at the putative disease locus. We consider dominant and recessive models as well as a family of generalized linear models (GLMs) that includes the additive, multiplicative, and linear logistic models. To use the test statistics based on the dominant and recessive models, we must specify the extent of gametic disequilibrium between marker and disease loci. For GLMs, however, the tests depend only on the total allele counts at the marker locus in affected and unaffected individuals. Thus, for these models, the tests can be used with pooled DNA.

Test Statistics

We assume that members in each of N unrelated families have been genotyped at a set of closely spaced markers in a chromosomal region. We also assume that, for some of the family members, phenotype (affected versus unaffected) is known. We want to use the marker data to test the null hypothesis that none of the genes in the region is related to the disease. Our objective is to develop test statistics with good power under the alternative hypothesis that a locus in the region is associated with the disease. We base our tests on the likelihood of the family's marker data, considered as a function of position t in the region. By use of the term "region," we define the set of all loci that are both linked to and in gametic disequilibrium with at least one of the markers. We make basic assumption A.1: given the family's genoytypes at a test locus t, the family's phenotypes and marker genotypes are independent. Figure 1 shows the chromosomal region when the test locus t lies between two markers.

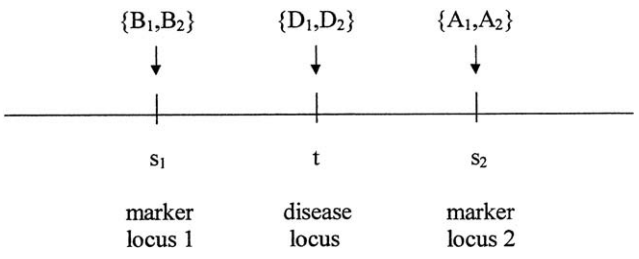

Figure 1 A portion of the chromosomal region of interest, in which the test locus t is flanked by two diallelic markers. The power of the test statistics is determined by (*a*) the extent of disequilibrium between t and the two markers among the chromosomes in the founder population and (*b*) the probabilities of recombination between t and each of the two markers in meiosis from parents to offspring within a family.

Likelihood for a Family

We define a family to be a set of individuals such that (*a*) any two individuals are connected in the sense proposed by Thompson (1986, p 21) and (*b*) each individual is either a founder (neither parent belongs to the family) or a nonfounder (both parents belong to the family). The genotype and/or phenotype of any family member may be unknown; however, the family must contain at least one member with a known phenotype and at least one member with a known genotype. Let $\mathbf{y} = (y_1,...,y_m)$ denote the vector of phenotypes for the m members with known phenotype. Here y_ℓ has a value of 1 if the ℓth family member is affected with the disease and has a value of 0 otherwise, $\ell = 1,...,m$. We assume that phenotypes are missing at random (Little and Rubin 1987)—that is, the probability of failure to observe a member's disease status does not depend on his or her actual status or on his or her marker genotypes. With this assumption, the likelihood at locus t is the probability $P(\mathcal{M}|R,\mathbf{y},t)$ of the family's observed marker data, denoted as $\mathcal{M}$, given the family's geneological structure R, given the vector $\mathbf{y}$ of observed phenotypes, and given that t is the disease locus. We use Bayes theorem to write this probability as follows:

$$P(\mathcal{M}|R,\mathbf{y},t) = \frac{P(\mathcal{M})\mathcal{P}(\mathbf{y}|\mathcal{M},R,t)}{P(\mathbf{y}|R,t)} \quad . \qquad (1)$$

To simplify the notation, we now suppress the dependence of the probabilities both on the family structure R and on the particular locus t. Let $\mathbf{g} = (g_1,...,g_m)$ denote the vector of genotypes at locus t of the m family members of known phenotype. We want to allow for the possibility that t is not one of the marker loci. Assumption A.1 states that the family's phenotype $\mathbf{y}$ and its

marker data $\mathcal{M}$ are conditionally independent, given **g**. Thus, we write

$$P(\mathbf{y}|\mathcal{M}) = \sum_{\mathbf{g}} P(\mathbf{y}|\mathbf{g})P(\mathbf{g}|\mathcal{M}) \ , \qquad (2)$$

where $\sum_{\mathbf{g}}$ denotes summation over all possible genotype vectors **g**. Substitution of notation (2) into probability (1) gives the likelihood for the family as

$$P(\mathcal{M}|\mathbf{y}) = P(\mathcal{M}) \sum_{\mathbf{g}} \frac{P(\mathbf{y}|\mathbf{g})}{P(\mathbf{y})} P(\mathbf{g}|\mathcal{M}) \ . \qquad (3)$$

In this instance, $P(\mathbf{y}) = \Sigma_{\mathbf{g}} P(\mathbf{y}|\mathbf{g})P(\mathbf{g})$ is the marginal probability of the family phenotype.

Family likelihood (3) involves three types of probabilities. The first type is the probability $P(\mathcal{M})$ of the family's marker data, under the null hypothesis that the chromosomal region containing the markers is not related to disease risk. We shall specify this probability in terms of a vector of *marker parameters*. These determine the frequencies of the marker alleles or haplotypes among chromosomes in the populations from which the family's founders were drawn, and we assume that they are known.

The second type of probability relates the alleles at locus t to the marker data. These probabilities appear in expression (3) as $P(\mathbf{g}|\mathcal{M})$. They depend on the extent of gametic disequilibrium between locus t and the marker loci in the family founders and on the probability of recombination between t and its nearest marker loci in meioses within a family. In general, the extent of gametic disequilibrium between markers and t is not known, unless t is one of the marker loci. However, in many situations of practical interest, the test statistics themselves do not depend on this gametic disequilibrium, although their power does.

The third type of probability needed in the likelihood consists of the penetrance functions $P(\mathbf{y}|\mathbf{g})$ for the various possible family genotypes **g**. For family members with known phenotype, these penetrance functions give the joint probability of disease occurrence or nonoccurrence as functions of the members' genotypes at locus t. To specify these penetrances, we assume that, at most, one allele or one group of alleles, labeled D_1, confers elevated disease risk. We also group all other alleles and label this group as allele D_2. We set $g = i$ for the genotype of an individual who carries i copies of the putative high-risk allele D_1, $i = 0,1,2$.

For families with one individual of known phenotype, we need only the three penetrances $P(y = 1|g = i)$, $i = 0,1,2$. We shall consider the following general class of penetrance models:

$$P(y = 1|g = i) = \pi(\alpha + \beta c_i) \ ,$$
$$i = 0,1,2; \text{ with } c_0 = 0, \ c_2 = 1 \ . \qquad (4)$$

Here π is a known smooth monotonic function, α is a constant that specifies risk in D_2D_2 homozygotes, and β is an unknown constant relating risk in D_1D_1 homozygotes to that in D_2D_2 homozygotes. Also, c_1 is a specified constant relating disease risk in D_1D_2 heterozygotes to that in D_2D_2 homozygotes. The value $\beta = 0$ corresponds to the null hypothesis of no relation between the disease and locus t. In this instance, the parameter α determines the disease prevalence in the population $\pi(\alpha)$, under the null hypothesis that disease is unrelated to the family genotype **g**. We will often assume that $\pi(\alpha)$ is known from data on disease prevalence in the population. We shall call α and β *"penetrance parameters."*

Penetrance has traditionally been modeled with the use of $\pi(x) = x$, with $c_1 = 1$ for the dominant model, $c_1 = 0$ for the recessive model, and $c_1 = 1/2$ for the additive model. Other models in the class (4) include the multiplicative model, with $\pi(x) = \exp(x)$ and $c_1 = 1/2$ (Self et al. 1991; Risch and Merikangas 1996; Schaid 1996; Whittaker and Lewis 1998), and the linear logistic model, with $\pi(x) = e^x/(1 + e^x)$ and $c_1 = 1/2$. The penetrances in any model (4) with $c_1 = 1/2$ are, after an appropriate transformation, linear in the number i of high-risk D_1 alleles. Accordingly, we shall call them "GLMs" (McCullagh and Nelder 1989). Individuals with genotype g at a putative trait locus t will be said to have a "D_1 *count* of c_g." Thus, if the penetrances are specified by a GLM, an individual's D_1 count is one-half of the number of his or her copies of allele D_1. In contrast, for a recessive or dominant model, an individual's D_1 count is 1 if he or she is a carrier of the high-risk genotype; otherwise, it is 0.

For families with $m > 1$ members of known phenotype, we must specify their joint probability of disease, conditional on their genotypes at locus t. We assume that, given his or her own genotype g, an individual's phenotype does not depend on the genotypes of his or her relatives and that, given **g**, the phenotypes of relatives are conditionally independent. Therefore,

$$P(\mathbf{y}|\mathbf{g}) = P(y_1,...,y_m|g_1,...,g_m)$$
$$= \prod_{\ell=1}^{m} [\pi(\alpha + \beta c_{g_\ell})]^{y_\ell} [1 - \pi(\alpha + \beta c_{g_\ell})]^{1-y_\ell} \ . \qquad (5)$$

These assumptions ignore the possibility that residual correlation in family phenotypes may be a result of other loci responsible for the disease or of shared, unmeasured risk factors. Fortuitously, the proposed statistical tests

remain valid (in the sense of having the correct null asymptotic p values) even if this residual correlation is ignored, and we shall ignore it hereafter. However, it is possible that more-accurate modeling of the correlation could improve statistical power, and this possibility requires investigation.

From (5), we see that, under the null hypothesis $\beta = 0$, the probability $P(\mathbf{y}|\mathbf{g})$ is independent of the family genotype $\mathbf{g}$; thus, the likelihood (3) is simply the null probability $P(\mathcal{M})$ of the marker data. Our objective is to derive efficient score statistics based on the likelihood (3) and to use them to test the null hypothesis $\beta = 0$ for various test loci t in the region covered by the markers. When assumption A.1 holds, the score statistics have standard Gaussian null distributions—asymptotically, as $N \to \infty$. After presentation of the statistics, we will discuss, in brief, the types of bias that can arise when assumption A.1 fails.

Score for a Family

We consider the use of likelihood-based efficient score statistics (Cox and Hinkley 1974) for testing of the null hypothesis that a family's phenotype $\mathbf{y}$ is independent of its genotype $\mathbf{g}$ at locus t—that is, $\beta = 0$. To describe these statistics, we suppose that there are K possible categories of marker genotypes for the family and that, if the marker genotypes were known for all members, the family could be classified in one and only one of the categories. For example, table 1 shows the $K = 15$ categories for a nuclear family with one offspring, when the marker data consist of genotypes at a single diallelic locus. In general, the family's marker category may not be known, because some members have not been typed at some or all of the marker loci. To deal with this uncertainty, we shall introduce a random variable x_k that represents the null probability that the family has category k, given its observed marker data $\mathcal{M}$:

$$x_k = P(\text{category} = k|\mathcal{M}) = \frac{r_k P(\mathcal{M}|\text{category} = k)}{\sum_{k'} r_{k'} P(\mathcal{M}|\text{category} = k')} \quad . \tag{6}$$

In this instance, r_k is the marginal probability that the family belongs to category k, under the null hypothesis. If the family is known to have, for example, category ℓ, then $x_\ell = 1$ and $x_k = 0$, $k \neq \ell$. If the category is not known, then x_k is a conditional probability, given the marker data.

In the Appendix, we show that, for the class (4) of penetrance models and on the basis of the likelihood (3) at locus t, the score for the family is

Table 1

***K* = 15 Categories of a Diallelic Autosomal Marker for a Nuclear Family with One Offspring**

Marker Category ($k = fh = f_1f_2h$)[a]	Probability[b] u_f	$v_{h\|f}$	$C_{1k}C_{2k}C_{3k}$[c]	w_k[d]
1. B_1B_1,B_1B_1,B_1B_1 (222)	u_2^2	1	111	$2 - \psi$
2. B_1B_1,B_1B_2,B_1B_1 (212)	u_1u_2	1/2	$1c_11$	$2 - \psi c_1$
3. B_1B_1,B_1B_2,B_1B_2 (211)	u_1u_2	1/2	$1c_1c_1$	$1 - \psi c_1 + c_1$
4. B_1B_2,B_1B_1,B_1B_1 (122)	u_1u_2	1/2	c_111	$c_1 - \psi + 1$
5. B_1B_2,B_1B_1,B_1B_2 (121)	u_1u_2	1/2	$c_1 1c_1$	$2c_1 - \psi$
6. B_1B_1,B_2B_2,B_1B_2 (201)	u_0u_2	1	$10c_1$	$1 + c_1$
7. B_2B_2,B_1B_1,B_1B_2 (021)	u_0u_2	1	$01c_1$	$-\psi + c_1$
8. B_1B_2,B_1B_2,B_1B_1 (112)	u_1^2	1/4	c_1c_11	$(1 - \psi)c_1 + 1$
9. B_1B_2,B_1B_2,B_1B_2 (111)	u_1^2	1/2	$c_1c_1c_1$	$(2 - \psi)c_1$
10. B_1B_2,B_1B_2,B_2B_2 (110)	u_1^2	1/4	c_1c_10	$(1 - \psi)c_1$
11. B_1B_2,B_2B_2,B_1B_2 (101)	u_1u_0	1/2	c_10c_1	$2c_1$
12. B_1B_2,B_2B_2,B_2B_2 (100)	u_1u_0	1/2	c_100	c_1
13. B_2B_2,B_1B_2,B_1B_2 (011)	u_0u_1	1/2	$0c_1c_1$	$(1 - \psi)c_1$
14. B_2B_2,B_1B_2,B_2B_2 (010)	u_0u_1	1/2	$0c_10$	$-\psi c_1$
15. B_2B_2,B_2B_2,B_2B_2 (000)	u_0^2	1	000	0

[a] f_i = number of B_1 alleles in the genotype of parent i, $i = 1,2$; h = number of B_1 alleles in genotype of the offspring.

[b] $u_f = u_{f_1f_2}$ = probability of parental genotype $f = f_1f_2$; $v_{h|f}$ = probability of offspring genotype h, given parental genotype f.

[c] $C_{\ell k} = D_1$ count of family member ℓ, $\ell = 1,2,3$, when family has category k and when $D_1 = B_1$.

[d] $w_k = a_1C_{1k} + a_2C_{2k} + a_3C_{3k}$, where a_ℓ is a phenotype value for family member ℓ, $\ell = 1,2,3$.

$$S = \sum_k w_k(x_k - r_k) \quad . \tag{7}$$

Here w_k is a weight attached to marker category k. Thus, the score is a weighted sum of deviations between "observed" and expected frequencies of the K marker categories. The weights w_k determine the importance of one marker category relative to another. They depend on the relations between genotypes at marker loci and genotypes at locus t as well as on the penetrance functions relating disease to genotypes at t.

To describe the weights, we consider the simple example of a nuclear family with one offspring evaluated at a single diallelic "candidate gene" with alleles $B_1 = D_1$ and $B_2 = D_2$ and with no missing genotypes or phenotypes. In table 1, the Marker Category column shows the 15 possible marker categories for the family, where f_1, f_2, and h denote the genotypes of parent 1, parent 2, and the offspring, respectively. For category $k = f_1f_2h$, the penetrance values c_2, c_1, and c_0 determine a triple value, $C_{1k}C_{2k}C_{3k} = c_{f_1}c_{f_2}c_h$, of D_1 counts for the family. The $C_{1k}C_{2k}C_{3k}$ column in table 1 shows these triple values for $c_2 = 1$ and $c_0 = 0$. The weight w_k for category k is

$$w_k = \sum_{\ell=1}^{n} a_\ell C_{\ell k} \ , \tag{8}$$

where $C_{\ell k}$ is the D_1 count of family member ℓ when the family has marker category k, $\ell = 1,...,n$, and $n = 3$ in this example. Also, a_ℓ, a phenotype value for member ℓ, is defined as follows:

$$a_\ell = \begin{cases} 1 & \text{if } y_\ell \text{ observed and } y_\ell = 1 \\ -\psi & \text{if } y_\ell \text{ observed and } y_\ell = 0 \\ 0 & \text{if } y_\ell \text{ unobserved .} \end{cases} \tag{9}$$

In this definition, $\psi = \pi/(1 - \pi)$, where $\pi = \pi(\alpha)$ is the disease prevalence in the population under the null hypothesis. If, for example, parent 1 and the offspring are affected and parent 2 is unaffected, then $a_1 = a_3 = 1$ and $a_2 = -\psi$. Substitution of these values in (8) results in $w_k = C_{1k} + C_{3k} - \psi C_{2k}$. Thus, w_k is a difference between D_1 counts of affected and unaffected family members, with unaffected family members contributing a value ψ, relative to affected members. In table 1, the w_k column shows the weights w_k for the 15 marker categories.

The marker category probability r_k in score (7) factors as the probability of the founder-genotype category multiplied by the conditional probability of the nonfounder-genotype category, given the founder-genotype category. Thus, a specific marker category can be written as $k = fh$, where f and h represent categories of marker genotypes for founders and nonfounders, respectively. For example, the family genotype B_1B_1,B_1B_2,B_1B_1 (table 1, row 3 [B_1B_1,B_1B_2,B_1B_2 {211}]) is labeled $f = 21$ and $h = 1$ or $k = fh = 211$. We write

$$r_k = r_{fh} = u_f v_{h|f} \ , \tag{10}$$

where u_f is the null probability of founder category f and where $v_{h|f}$ is the probability of nonfounder subcategory h, given that the founders' genotypes belong to category f. The Probability column of table 1 shows u_f and $v_{h|f}$ for the nuclear family in this example. Notice that the probabilities r_k depend on the marker parameters only through u_f; the probabilities $v_{h|f}$ are constants determined by the Mendelian laws of inheritance.

To illustrate computation of the score S, suppose that the nuclear family is observed to have marker category $f_1f_2h = 212$. Then, from equation (6), $x_{212} = 1$ and $x_k = 0$, $k \neq 212$. Substitution of these values into equation (7) gives the score for this family as $S = w_{212} - \Sigma_k r_k w_k$. From equation (8), we see that the score is the difference between the observed and expected values of a linear combination of D_1 counts among the family members.

In row 2 (B_1B_1,B_1B_2,B_1B_1 [212]) of table 1, $w_{212} = 2 - \psi c_1$. Also, from equation (8) and from the values for r_k in table 1, $\Sigma_k r_k w_k = \Sigma_{\ell=1}^{3} a_\ell(\Sigma_k r_k C_{\ell k}) = (a_1 + a_2)(u_2 + c_1u_1) + a_3(u'_2 + c_1u'_1)$, where $u'_2 = (u_2 + \frac{1}{2}u_1)^2$ and $u'_1 = u_1 + 2u_0u_2 - \frac{1}{2}u_1^2$. Thus, $S = 2 - \psi c_1 - [(1 - \psi)(u_2 + c_1u_1) + u'_2 + c_1u'_1]$. Under Hardy-Weinberg equilibrium (HWE) for the parental-genotype frequencies, $u_2 = u'_2 = [P(B_1)]^2$ and $u_1 = u'_1 = 2P(B_1)P(B_2)$. If, in addition, $c_1 = 1/2$, then $S = 2 - \frac{1}{2}\psi - (2 - \psi)P(B_1)$.

Suppose that parent 1 is untyped. We therefore know only that the family marker category is either 212 or 112. Thus, according to equation (6) and rows 2 and 8 (B_1B_1,B_1B_2,B_1B_1 [212] and B_1B_2,B_1B_2,B_1B_1 [112], respectively) of table 1,

$$x_{212} = \frac{r_{212}}{r_{212} + r_{112}} = \frac{2u_2}{2u_2 + u_1}$$

$$x_{112} = \frac{r_{112}}{r_{212} + r_{112}} = \frac{u_1}{2u_2 + u_1}$$

$$x_k = 0, \ k \neq 212, \ 112 \ .$$

Under HWE, $x_{212} = P(B_1)$ and $x_{112} = P(B_2)$. The family's score is as follows:

$$\begin{aligned} S &= \frac{2u_2}{2u_2 + u_1} w_{212} + \frac{u_1}{2u_2 + u_1} w_{112} - \sum_k r_k w_k \\ &= \frac{2u_2}{2u_2 + u_1}(2 - \psi c_1) + \frac{u_1}{2u_2 + u_1}(1 - \psi c_1 + c_1) \\ &\quad - \sum_k r_k w_k \ . \end{aligned}$$

In general, when some family members are untyped, a random vector of probabilities (6) is assigned to the possible categories, where the probabilities are conditional on the observed marker data for the family. This random vector depends on the null founder-genotype probabilities u_f, which must be specified or estimated from external data. In the companion article (Tu et al. 2000), we illustrate maximum-likelihood estimation of u_f when the founders are parents in nuclear families. Martin et al. (1998) have also proposed such genotype reconstruction for parents in nuclear families.

We will now describe the weights w_k for families of arbitrary structure, not only when some members may be untyped but, also, when the markers do not necessarily include the trait locus. To do so, we will let $\gamma_{\ell k}(i)$ denote the probability that member ℓ of a family with marker category k carries i copies of allele D_1, $i = 0,1,2$. The weights are then given according to equation (8), where $C_{\ell k} = c_1\gamma_{\ell k}(1) + \gamma_{\ell k}(2)$ is now the conditional D_1 count for family member ℓ, given the family marker category k. When, as in the example shown in table 1, the marker data include genotypes at locus t, a family marker category k specifies a D_1 count for each family member.

In summary, computation of a family's score S may

involve two types of imputation: (*a*) imputation of the family marker category when it is incompletely observed (e.g., when parental genotypes are missing) and (*b*) imputation of family members' D_1 counts at the putative disease locus t, when locus t is not one of the markers. The latter imputation requires knowledge of the extent of disequilibrium between locus t and its nearest markers among the family founders. Like the disease prevalence π, disequilibrium parameters relating alleles at locus t to those at the marker loci cannot be estimated from the data, since genotypes at locus t are not observed. However, misspecification of the disequilibrium parameters will not affect the validity of a score test, although it will decrease the test's power.

Expressions (8) and (9) indicate that the contribution of an unaffected member, relative to that of an affected member, is determined by the disease odds $\psi = \pi/(1 - \pi)$. Since the disease prevalence π is invariably <1/2, the counts of unaffected members contribute less to the score than do those of affected members. In particular, for rare diseases ($\pi \ll 1$), the counts of unaffected members contribute little to the score statistic. For many diseases, π is known at least approximately. It cannot be estimated from the family data, since the families have been ascertained on the basis of their phenotype. However, since π appears only in the weights w_k and not in the null probabilities r_k, its misspecification will not produce incorrect p values, although it may affect statistical power. Choosing optimal values for ψ is an area that requires research. In the companion article (Tu et al. 2000), we use simulations to compare the power of tests based on genotypes in phenotype-discordant sib pairs, with the use of $\psi = 1$ and ψ equal to the disease odds in the population.

Under the null hypothesis, the score S of equation (7) has a mean of 0 and an asymptotic variance of

$$V = E[S^2] = E\left[\left(\sum_k w_k x_k\right)^2\right] - \left(\sum_k w_k r_k\right)^2 , \qquad (11)$$

as described in the Appendix.

Decomposition of the Score

The factorization (10) of the null family marker category probabilities r_k of induces a decomposition of the score (7) as a sum of nonfounder and founder scores: $S = S_{NF} + S_F$. In this instance, the nonfounder score is

$$S_{NF} = \sum_f \sum_{h|f} w_{fh}(x_{fh} - x_{f\cdot} v_{h|f}) , \qquad (12)$$

where $\sum_{h|f}$ denotes summation over the nonfounder categories h compatible with f, $w_{fh} = w_k$, and $x_{fh} = x_k$ when $fh = k$, and with $x_{f\cdot} = \sum_{h|f} x_{fh}$.

For the family with observed marker category 212 (see table 1), we have $x_{21\cdot} = 1$ and $x_{f\cdot} = 0$ when $f \neq 21$. Substitution of these values and the corresponding $v_{h|f}$ into equation (12) results in the following nonfounder score for this family: $S_{NF} = w_{212} - \frac{1}{2}(w_{212} + w_{211}) = \frac{1}{2}(w_{212} - w_{211})$. For the phenotype values $(a_1,a_2,a_3) = (1, -\psi, 1)$, w_{212} and w_{211} are shown in rows 2 and 3 (B_1B_1,B_1B_2,B_1B_1 [212] and B_1B_1,B_1B_2,B_1B_2 [211], respectively) of table 1. With these values, $S_{NF} = 2 - \psi c_1 - \frac{1}{2}(3 - 2\psi c_1 + c_1) = \frac{1}{2}(1 - c_1)$. Marker category 212 denotes transmission of allele B_1 from heterozygous parent 2 to the affected offspring. If this parent had transmitted allele B_2 to the affected offspring, resulting in marker category 211, then the NFS would be $S_{NF} = \frac{1}{2}(c_1 - 1)$. Thus, if $c_1 = 1/2$, then $S_{NF} = 1/4$ or $-1/4$, depending on whether parent 2 transmits allele B_1 or B_2 to the offspring. Indeed, for a single marker in nuclear families with known genotypes and $c_1 = 1/2$, S_{NF} is just the TDT.

Notice that the parental-phenotype values do not contribute to S_{NF}. In general, founders contribute only their genotypes to the NFS, and these founder genotypes are used only to compute the null expectations of the nonfounder genotypes. In contrast, phenotypes of nonfounders play a central role in the statistic. If, for example, the phenotype of the single offspring in this family were unknown, then S_{NF} would vanish. The genotypes of nonfounders with unknown phenotype are useful for reconstruction of the family's marker category, but, because they do not enter the weights w_k, they are not evaluated directly in the NFS.

When the genotype of parent 1 is unknown, the nonfounder score is as follows:

$$S_{NF} = \frac{2u_2}{2u_2 + u_1}(2 - \psi c_1) + \frac{u_1}{2u_2 + u_1}(1 - \psi c_1 + c_1) - \frac{1}{2}(3 - 2\psi c_1 - c_1) .$$

As this example illustrates, when the founder category of the family is known, S_{NF} does not involve the marker parameters. However, when the founder category is not known, S_{NF} depends on the marker parameters through the founder-category probabilities u_f, which, in formula (6), determine the r_k for the x_k.

The variance of S_{NF}, conditional on the vector $\mathbf{x}_\cdot = (x_{1\cdot},...,x_{F\cdot})$ of probabilities for the F founder categories, is as follows:

$$V_{NF} = \sum_f x_{f\cdot}\left\{E\left[\left(\sum_{h|f} w_{fh} x_{fh}\right)^2\right] - \left(\sum_{h|f} w_{fh} v_{h|f}\right)^2\right\} .$$

We now turn to the founder score, which is as follows:

$$S_{\mathrm{F}} = \sum_{f} \overline{w}_f (x_{f\cdot} - u_f) \ . \tag{14}$$

In this case,

$$\overline{w}_f = \sum_{h|f} w_{fh} v_{h|f} = \sum_{\ell=1}^{n} a_\ell \overline{C}_{\ell f} \ , \tag{15}$$

and $\overline{C}_{\ell f} = \Sigma_{h|f} C_{\ell fh} v_{h|f}$ is the expected D_1 count for the ℓth member in a family whose founders have marker category f. Consider again the family with category 212 (see table 1), so that $x_{21\cdot} = 1$ and $x_{f\cdot} = 0$ when $f \neq 21$. For this family, $S_{\mathrm{F}} = \overline{w}_{21} - \Sigma_{f_1=0}^{2} \Sigma_{f_2=0}^{2} \overline{w}_{f_1 f_2} u_{f_1} u_{f_2}$. From equation (15), we see that, with phenotype values $(a_1, a_2, a_3) = (1, -\psi, 1)$,

$$\overline{w}_{21} = \overline{C}_{1,21} - \psi \overline{C}_{2,21} + \overline{C}_{3,21} = \frac{3}{2} - \psi c_1 + \frac{1}{2} c_1 \ ,$$

where we have used the expected D_1 counts $\overline{C}_{1,21} = 1$, $\overline{C}_{2,21} = c_1$, and $\overline{C}_{3,21} = \frac{1}{2}(1 + c_1)$. Two observations are noteworthy in this case: (*a*) the expected D_1 counts for the founder parents are just their observed D_1 counts, whereas the D_1 count of the offspring is his or her expected count, given those of his or her parents, and (*b*) if, for example, parent 1 has a known genotype but an unknown phenotype—so that $a_1 = 0$—then the genotype of this parent still contributes to the founder score through its contribution to the expected D_1 count of his or her affected offspring. In contrast, if the offspring has a known genotype but an unknown phenotype, his or her genotype is used only to reconstruct the genotypes of his or her parents.

The variance of S_{F} is as follows:

$$V_{\mathrm{F}} = E[S_{\mathrm{F}}^2] = E\left[\left(\sum_f \overline{w}_f x_{f\cdot}\right)^2\right] - \left(\sum_f \overline{w}_f u_f\right)^2 \ . \tag{16}$$

The variance of the total score, given by equation (11), can be written as follows:

$$V = V_{\mathrm{F}} + E[V_{\mathrm{NF}}] \ , \tag{17}$$

where $E[V_{\mathrm{NF}}]$ is given by equation (13), with $x_{f\cdot}$ replaced with u_f.

Score Statistic for N Families

For a collection of N independent families from a population that is homogeneous with respect to disease risk, the nonfounder, founder, and total score statistics are, respectively,

$$\mathcal{T}_{\mathrm{NF}} = \frac{\sum_{\nu=1}^{N} S_{\mathrm{NF}\nu}}{\sqrt{\sum_\nu V_{\mathrm{NF}\nu}}} \ , \ \mathcal{T}_{\mathrm{F}} = \frac{\sum_{\nu=1}^{N} S_{\mathrm{F}\nu}}{\sqrt{\sum_\nu \hat{V}_{\mathrm{F}\nu}}} \ ,$$

$$\text{and } \mathcal{T} = \frac{\sum_{\nu=1}^{N} (S_{\mathrm{NF}\nu} + S_{\mathrm{F}\nu})}{\sqrt{\sum_\nu \hat{V}_\nu}} \ . \tag{18}$$

In this case, for the νth family, $\nu = 1,\ldots,N$, $V_{\mathrm{NF}\nu}$, $V_{\mathrm{F}\nu}$, and V_ν, are given by (13), (16), and (17), and the hat denotes an estimate. (In these expressions, the unknown parameters u_f must be replaced with estimates.) When the founder-genotype probabilities are specified correctly, the statistics $\mathcal{T}$, $\mathcal{T}_{\mathrm{NF}}$, and $\mathcal{T}_{\mathrm{F}}$ in (18) each have—asymptotically, as $N \to \infty$—a standard Gaussian distribution under the null hypothesis of no effect at locus t.

The score statistics for N families from a population that is heterogeneous with respect to disease risk are the weighted averages of statistics for the I homogeneous subpopulations. For example, the total score statistic is $\mathcal{T} = \Sigma_{i=1}^{I} \epsilon_i T_i / \sqrt{\Sigma_{i=1}^{I} \epsilon_i^2}$, where, for subpopulation i, $\mathcal{T}_i$ is the total score statistic and where the weight ϵ_i depends on its disease risk in the population π_i, as described in the Appendix.

Some invariance properties of the score statistics are worth noting. First, a family's founder score $S_{\mathrm{F}\nu}$ in (14) is unchanged if all weights $\overline{w}_{\nu f}$ are replaced with $\overline{w}^*_{\nu f} = \overline{w}_{\nu f} + \xi_\nu$, where ξ_ν is an arbitrary family-specific constant, since $\sum_f x_{\nu f} = \sum_f u_f = 1$. Second, the nonfounder score $S_{\mathrm{NF}\nu}$ in (12) is unchanged when all weights $w_{\nu fh}$ are replaced by $w^*_{\nu fh} = w_{\nu fh} + \xi_{\nu f}$, where $\xi_{\nu f}$ is an arbitrary family-specific constant that is independent of its nonfounder category h. In particular, since the conditional D_1 counts of founders depend only on their own category f and not on the category h of their descendants, the summands of $w_{\nu fh}$ corresponding to founders can be ignored. Therefore, neither the D_1 counts nor the phenotypes of founders contribute to the NFS. Finally, the weights in either the founder score or the nonfounder score can all be multiplied by any nonzero constant ξ, without changing the standardized test statistics. We shall use these invariance properties to simplify the test statistics in the examples in this article and in the companion article (Tu et al. 2000).

We conclude this section with a brief discussion of the bias (i.e., incorrect asymptotic type I–error probabilities) that could arise if assumption A.1 fails. This assumption states that, given the family's genotypes at a test locus t, its phenotypes and marker genotypes are independent. It would fail if one or more of the markers were asso-

ciated with any (genetic or nongenetic) risk factors for the disease. In this case, the likelihood at locus *t,* under the null hypothesis $\beta = 0$, would not be the null probability of the markers in the general population, since the families were ascertained for multiple cases of disease. Consequently, the FS, which compares the genotype frequencies among founders in these multiple-case families with those in some reference population, would not have a standard Gaussian distribution. Thus, the FS can be biased by association between markers and risk factors that do not segregate with disease within families. The NFS, in contrast, is conditioned on the observed or inferred distribution of marker genotypes in the founders of the ascertained families. This conditioning assures that the NFS has the correct asymptotic null distribution in the absence of linkage to a locus that segregates with disease within families, provided that, when some founder genotypes are unobserved, the distribution of founder genotypes is specified correctly. While departures from random mating, HWE, etc., in the founders could, in principle, affect the asymptotic null distribution of the NFS, it is difficult to envisage examples involving serious bias. Most likely, divergent results for NFS and FS at a given locus *t,* with the former being nonsignificant and the latter significant, would suggest marker-disease association in the absence of linkage. Conversely, if the NFS were significant but the FS were nonsignificant, then this would suggest that the markers are linked to, but are in weak gametic disequilibrium with, a disease locus *t.*

Application to Single Diallelic Polymorphisms in Case Series and Case-Control Studies

We illustrate the score statistics by applying them to very simple families, such as nuclear families and "families" consisting of single unrelated individuals. We regard a single individual, who is either affected (a case) or unaffected (a control), as a founder of his or her "family." The scores of nuclear families and of unrelated cases and controls are summed to form the test statistics. The FS T_F evaluates the genotypes of cases, controls, and parents in the nuclear families, whereas the NSF T_{NF} evaluates just the genotypes of the offspring in the nuclear families, comparing them with the Mendelian expectation.

Suppose that individuals are typed at a single diallelic marker with alleles B_1 and B_2. The marker locus may be distinct from the test locus *t,* but, if this is so, then the alleles at the two loci are assumed to be in gametic disequilibrium in the population. Let $P(D_iB_j)$ denote the probability that a random chromosome from the population carries the haplotype D_iB_j, $i, j = 1, 2$. Table 2 gives these probabilities in terms of the marginal probabilities $P(D_i)$ and $P(B_j)$ and the disequilibrium coefficient $\delta = P(D_1B_1) - P(D_1)P(B_1)$. The probability p_1 that a random chromosome containing allele B_1 also contains allele D_1 is thus

$$p_1 = 1 - q_1 = P(D_1|B_1) = \frac{P(D_1B_1)}{P(B_1)} = P(D_1) + \frac{\delta}{P(B_1)} \ .$$

The analogous probability p_2, for a chromosome containing B_2, is

$$p_2 = 1 - q_2 = P(D_1|B_2) = P(D_1) - \frac{\delta}{P(B_2)} \ .$$

For convenience, we introduce a standardized disequilibrium coefficient Δ, defined as $\Delta = \delta/[P(B_1)P(B_2)] = p_1 - p_2$. By assumption, $\Delta \neq 0$.

In the companion article (Tu et al. 2000), we applied the score statistics to nuclear families for which markers at the putative disease locus are observed but for which the family marker category typically is unobserved. In contrast, in the present study, we apply them to families consisting of single, unrelated individuals in which the marker category is observed but in which the marker may not be the trait locus. Since these latter families contain only founders, the nonfounder score is 0, and the total score for an individual reduces to the founder score S_F of (14). There are $K = F = 3$ marker categories—namely, the three genotypes B_2B_2, B_1B_2, and B_1B_1—which are indexed as $f = 0,1,2$, respectively. Their null probabilities are $r_f = u_f$, $f = 0,1,2$.

Case Series

In case series, the marker genotypes of a sample of N cases are compared with those in some reference population. The score statistic is given by (14), with the weight for a case with genotype f given as $w_f = \overline{w}_f = \overline{C}_f$, as shown in table 3. The invariance properties of the test statistic, which were described at the end of the previous section, allow us to standardize the weights so that $w_0 = 0$ and $w_2 = 1$. Note, however, that this standardization requires replacement of $w_f = \overline{C}_f$ with $w_f =$

Table 2

Haplotype Probabilities at Test Locus *t* and Marker Locus *s*

	Haplotype Probability at Marker Allele		
Disease Allele	B_1	B_2	Total
D_1	$P(B_1)P(D_1) + \delta$	$P(B_2)P(D_1) - \delta$	$P(D_1)$
D_2	$P(B_1)P(D_2) - \delta$	$P(B_2)P(D_2) + \delta$	$P(D_2)$
Total	$P(B_1)$	$P(B_2)$	1

Table 3

Probability $\overline{\gamma}_f(i)$ That a Founder with f Copies of Marker Allele B_1 Carries i Copies of a Nearby Disease Allele

f	$\overline{\gamma}_f(0)$	$\overline{\gamma}_f(1)$	$\overline{\gamma}_f(2)$	$\overline{C}_f = c_1\overline{\gamma}_f(1) + \overline{\gamma}_f(2)$	$\frac{\overline{C}_f - \overline{C}_0}{\overline{C}_2 - \overline{C}_0}$ [a]
2	q_1^2	$2p_1q_1$	p_1^2	$2c_1p_1 + ep_1^2$ [b]	1
1	q_1q_2	$p_1q_2 + q_1p_2$	p_1p_2	$c_1(p_1 + p_2) + ep_1p_2$	λ [c]
0	q_2^2	$2p_2q_2$	p_2^2	$2c_1p_2 + ep_2^2$	0

[a] $\overline{C}_2 - \overline{C}_0 = \Delta[2c_1 + (1 - 2c_1)(p_1 + p_2)] \neq 0$, if and only if $\Delta \neq 0$, where $\Delta = p_1 - p_2$ is the standardized disequilibrium coefficient.
[b] $e = 1 - 2c_1$.
[c] $\lambda = (c_1 + ep_2)/[2c_1 + e(p_1 + p_2)]$.

$(\overline{C}_f - \overline{C}_0)/(\overline{C}_2 - \overline{C}_0)$, where $\overline{C}_2 - \overline{C}_0 = \Delta[2c_1 + (1 + 2c_1)(p_1 + p_2)]$. Since $\overline{C}_2 - \overline{C}_0$ is proportional to the disequilibrium coefficient Δ, the standardization is possible only when $\Delta \neq 0$. Indeed, S_F is itself proportional to Δ. Thus, if $\Delta = 0$—that is, if the marker were in linkage equilibrium with the test locus, then the distribution of the FS would be degenerate at 0.

In table 3, we see that the weight attached to B_1B_2 heterozygotes, when B_1B_1 and B_2B_2 homozygotes receive weights of 1 and 0, respectively, is λ, which is defined as follows:

$$w_1 = \frac{c_1 + ep_2}{2c_1 + e(p_1 + p_2)} \equiv \lambda \ . \qquad (19)$$

Here $e = 1 - 2c_1$. We use the symbol $S_{F\nu}$ to denote the founder score for the νth case. Substitution of the weights (19) into equation (14) results in

$$S_{F\nu} = x_{\nu 2} + \lambda x_{\nu 1} - (u_2 + \lambda u_1) \ , \qquad (20)$$

where u_f is the frequency of genotype f in the reference population and where $x_{\nu f} = 1$; otherwise, if the case's genotype is f, $x_{\nu f} = 0$, $f = 0,1,2$. Summing the individual scores (20) over the N cases shows that the score is $\Sigma_{\nu=1}^{N} S_{F\nu} = N_2 + \lambda N_1 - N(u_2 + \lambda u_1)$, where N_f is the number of cases with genotype f. The founder (and total) score statistic is

$$T = \frac{N_2 + \lambda N_1 - N(u_2 + \lambda u_1)}{\sqrt{N}\sigma} \ , \qquad (21)$$

where, from (16),

$$\sigma^2 = V_F = u_2(1 - u_2) + \lambda^2 u_1(1 - u_1) - 2\lambda u_1 u_2 \ . \qquad (22)$$

To use T, we must specify the weight λ for heterozygotes and the genotype frequencies u_0,u_1,u_2 in the reference population.

Case-Control Studies

Case-control studies are based on comparison of the marker genotypes of, for example, N_ℓ cases with those of N_u controls. The likelihood for the data is the product of the probabilities of case and control marker data, and the corresponding score is the sum of the score for the case data plus the score for the control data. The resulting test statistic can be written as

$$T = \frac{T_\ell \psi \phi T_u}{\sqrt{1 + \psi^2\phi^2}} \ , \qquad (23)$$

where T_ℓ and T_u are the test statistics (21) applied to cases and controls, respectively; where ψ is the phenotype value for controls in equation (9); and where $\phi^2 = N_u/N_\ell$ is the control:case ratio.

To use T, we must specify the heterozygote weight λ of equation (19), the phenotype value ψ, and the null marker genotype probabilities u_0,u_1,u_2. With the choice of $\psi = N_\ell/N_u$, equation (23) becomes

$$T = \frac{[(N_{\ell 2} + \lambda N_{\ell 1})/N_\ell] - [(N_{u2} + \lambda N_{u1})/N_u]}{\sigma\sqrt{(1/N_\ell) + (1/N_u)}} \ ,$$

where $N_{\ell g}$ and N_{ug} are the numbers of cases and controls, respectively, with genotype $g = 0,1,2$ and where σ^2 is given by (22). This choice for ψ eliminates the null genotype frequencies u_0,u_1,u_2 from the numerator of T; however, they appear in σ. They can be estimated from the group of controls as $\hat{u}_f = (N_{uf})/(N_u)$, $f = 0,1,2$. The resulting test statistic T is the standardized difference in expected D_1 counts between cases and controls. For $c_1 = 1/2$, T is the statistic proposed by Barcellos et al. (1997) and Risch and Teng (1998), for testing of allelic association between disease and marker in pooled DNA from cases and controls. In this case, T also is the score statistic for the traditional linear logistic regression of unrelated cases and controls.

Heterozygote Weight λ

The optimal weight λ for heterozygotes varies according to the penetrance model and the extent of disequilibrium between trait and marker loci, as specified by the probabilities p_1 and p_2. In practice, these probabilities seldom are known. However, it is evident from (19) that, for $c_1 = 1/2$, the weight is $\lambda = 1/2$, independent of p_1 and p_2. For these models, T is just the standardized difference between observed and expected B_1 counts, regardless of the extent of gametic disequilibrium between the two loci (provided, of course, that some disequilibrium exists).

When trait and marker loci coincide (so that, for example, $p_2 = 0$, $p_1 = 1$), then $\lambda = c_1$. In this case, T is the

standardized difference between observed and expected B_1 counts in cases and controls. If trait and marker loci do not coincide and if $c_1 \neq \frac{1}{2}$, then the extent of disequilibrium between trait and marker loci can have a large influence on the heterozygote weight λ. For a dominant model ($c_1 = 1$), equation (19) shows that $\lambda = q_2/(q_1 + q_2)$, which is closer to 1/2 than to 1 when the disease allele is rare. For a recessive model ($c_1 = 0$), equation (19) shows $\lambda = p_2/(p_1 + p_2)$, which varies from 0 (when the disease allele and allele B_1 are in complete disequilibrium) to 1 (when the disease allele and allele B_2 are in complete disequilibrium).

Suppose, for example, that the frequency of the disease allele is .01 and that the two marker alleles are equally likely. Then, for maximum disequilibrium between disease and marker loci, $\lambda = 0$ for a recessive disease gene and $\lambda = .505$ for a dominant gene. If the disequilibrium coefficient δ equals one-fourth of its maximum value, then $\lambda = .25$ for a recessive gene and .503 for a dominant gene. Thus, when disease and marker loci do not coincide, the optimal weight λ for heterozygotes is not 0 for a recessive gene and 1 for a dominant gene. For rare disease alleles, it is always ~1/2, when the disease allele is dominant. For common disease alleles, the optimal weight can strongly depend on the tightness of disequilibrium, regardless of the mode of inheritance. For instance, when the disease-allele frequency is .2, then, under maximum disequilibrium, $\lambda = 0$ for a recessive model and $\lambda = .63$ for a dominant model. However, for loose disequilibrium (δ equal to one-fourth of its maximum value), $\lambda = .38$ for a recessive model and $\lambda = .53$ for a dominant model. Lack of knowledge concerning both the extent of gametic disequilibrium and the correct penetrance model suggests use of the value $\lambda = 1/2$ in the test statistic (21). There is a need for investigation of the power associated with use of this strategy, for a range of situations.

Discussion

We have described a likelihood-based score statistic for detection of disease genes by evaluation of phenotype-marker associations and transmission disequilibrium within families. The score statistic decomposes into two components, the NFS and the FS. These components represent two types of deviation between observed genotypes and their expectations under the null hypothesis. Each will have a large absolute value if the chromosomal region contains a disease locus and if the families have been selected to contain affected individuals. The NFS represents the deviation between observed and expected marker alleles in nonfounders, given the founder genotypes. It reflects transmission disequilibrium from parent to affected and unaffected offspring, and it will be large because the disease locus is linked to the markers. When the marker data consist of genotypes at a single locus, both the nonfounder and founder scores are proportional to the standardized disequilibrium coefficient Δ between marker and disease loci. Thus, the statistics should be used only to evaluate the etiologic relevance of loci t that are in linkage disequilibrium with the marker. In addition, the nonfounder score is proportional to $1 - 2\theta$, where θ is the probability of recombination between the marker and disease loci. Thus, if the two loci were unlinked ($\theta = 1/2$), then the nonfounder score would vanish identically, and the marker genotypes of nonfounders would not contribute to the total test statistic. If, on the other hand, $\theta = 0$, which would hold when we wish to test the etiologic relevance of the marker itself, then founders and nonfounders contribute equally to the total test statistic. In this sense, the genetic distance between marker and disease loci, as measured by their recombination fraction θ, determines the relative contributions of founder and nonfounder genotypes to the total test statistic.

The FS evaluates association between disease and marker alleles in family founders, and it extends the test statistics that are currently used for the analysis of case-control data. FS reflects deviation between the observed or inferred frequencies of marker genotypes in the founders and those that are expected in the general populations to which they belong. It measures association between the disease and the disease locus, and it will be large when there is gametic disequilibrium between disease and marker loci among the founder chromosomes. If the null founder-genotype frequencies can be estimated from independent data, then the FS provides information supporting or refuting the null hypothesis derived from the (observed or inferred) marker genotypes of the founders, even when their phenotypes are unknown. However, the FS also can be large because of population stratification or inappropriate assumptions (e.g., random mating or Hardy-Weinberg proportions) on the distribution of founder genotypes. Thus, the FS tests for association, whereas the NFS tests for linkage as the cause of the association.

These likelihood-based statistics have certain strengths and limitations. Because they are model-based, they clarify the role of the underlying genetic model in the determination of the weights for affected versus unaffected individuals and, when dealing with single diallelic markers, the weights for heterozygotes versus homozygotes. When the model is correctly specified, the statistics enjoy certain local asymptotic optimality properties (Cox and Hinkley 1974). However, the model requires assumptions about the distribution of the founder genotypes, and the effects on bias and power resulting from departures from these assumptions have not yet been fully evaluated. There is a need to assess the impact on bias and power loss

associated with misspecification of the distribution of founder genotypes. One might expect that this impact would be small when the founder genotypes are known or can be inferred with some confidence.

One strength of the statistics is that they apply to any type of family structure, including a "family" consisting of a single individual, and that they thus eliminate the need for many different ad hoc tests. In addition, the approach provides a strategy for dealing with missing phenotypes and genotypes for key family members, such as parents. Also, when it is applied to families in which the phenotypes of some founders are known, the FS allows comparison of genotypes of affected and unaffected founders. Simultaneous evaluation of the FS (which may be biased by population stratification) and the NFS (which is less vulnerable to such bias) can provide insight into the etiologic relevance of observed associations.

The likelihood-based framework presented in this study stimulates consideration of several potentially useful extensions. First, the likelihood could be extended to accommodate censored survival data rather than binary disease outcomes. Second, the likelihood could be modified to include nongenetic covariates, in the manner considered by Self et al. (1991). In fact, the likelihood function proposed by Self et al. is a special case of the nonfounder component of the likelihood considered in the present study. Inclusion of nongenetic covariates would lead to score statistics that have been adjusted for the effects of the covariates. In addition, joint maximization of the likelihood, with respect to regression coefficients for both genetic markers and nongenetic factors, would allow for multivariate estimation of genotype relative risks (Schaid and Sommer 1993; Schaid and Li 1997; Witte et al. 1999).

If it becomes feasible to produce reliable estimations of both intermarker genetic distances and population-specific intermarker disequilibrium coefficients, then association studies will benefit from simultaneous consideration of multiple markers that may flank a disease locus. The systematic framework presented here should prove useful for such studies.

Note added in proof.—Further discussion of likelihood-based methods analogous to the methods presented here can be found in a study by Clayton (1999).

Acknowledgment

This research was supported by National Institutes of Health grant R35-CA47448. The authors thank Joseph B. Keller and the reviewers, for helpful comments on an earlier version of this manuscript.

Appendix

We derive the score statistic for a family with phenotype $\mathbf{y} = (y_1,...,y_m)$, where m is the number of members with known phenotype. Suppose that there are K categories of marker genotypes. Let r_k denote the family's null probability of having category k, and let $x_k = P(k|\mathcal{M})$ denote its conditional probability of having category k, given its observed marker data $\mathcal{M}$. The likelihood (3) for the family can therefore be written as follows:

$$L(\Theta) = P(\mathcal{M}|\mathbf{y}) = \frac{P(\mathcal{M}) \sum_k x_k \sum_{\mathbf{g}} P(\mathbf{g}|k) P(\mathbf{y}|\mathbf{g};\alpha,\beta)}{\sum_{\mathbf{g}} P(\mathbf{g}) P(\mathbf{y}|\mathbf{g};\alpha,\beta)} \quad . \tag{A1}$$

$P(\mathbf{g}|k)$ is the probability that a family with marker category k has genotype $\mathbf{g}$ at the disease locus t. Also, Θ is a vector of parameters that includes the penetrance parameters α and β, any unknown marker parameters in the probabilities r_k, and the test-locus-vs.-marker parameters. Let $\tilde{\Theta}$ be a null value of Θ—that is, one for which $\beta = 0$ and for which the remaining parameters are specified under the null hypothesis. By differentiation of the logarithm of (A1), with respect to β, and by evaluation of the same logarithm at $\tilde{\Theta}$, we find, after some algebraic calculations, that the family's score is as follows:

$$\frac{\partial}{\partial \beta} \log L(\Theta)|_{\Theta=\tilde{\Theta}} = \epsilon \sum_k w_k (x_k - r_k) \equiv \epsilon S \ .$$

In this instance, $\epsilon = \frac{d}{d\alpha} \log \pi(\alpha)|_{\alpha=\tilde{\alpha}}$ is the logarithmic derivative of the null disease prevalence in the population, and w_k is a nonnegative constant, as described in equations (8) and (9).

The null mean of S is 0, which follows from likelihood theory (Cox and Hinkley 1974). This can also be seen from equation (7) and from the fact that the null mean of the random variable

$$x_k = x_k(\mathcal{M}) = \frac{r_k P(\mathcal{M}|\text{category} = k)}{P(\mathcal{M})} \tag{A2}$$

is

$$E[x_k(\mathcal{M})] = \sum_{\mathcal{M}} P(\mathcal{M}) x_k(\mathcal{M}) = r_k \sum_{\mathcal{M}} P(\mathcal{M}|\text{category} = k) = r_k \ .$$

$\sum_{\mathcal{M}}$ denotes summation over all possible realizations of the observed marker data $\mathcal{M}$.

The asymptotic variance of the score ϵS (Cox and Hinkley 1974) is

$$E\left[-\frac{\partial^2 \log L(\Theta)}{\partial^2 \beta}\,|_{\Theta=\tilde{\Theta}}\right] = \epsilon^2\left\{E\left[\left(\sum_k w_k x_k\right)^2\right] - \left(\sum_k w_k r_k\right)^2\right\} \equiv \epsilon^2 V \ .$$

For N families from a population that is homogeneous with respect to disease risk π, the score statistic is $\mathcal{T} = \sum_{\nu=1}^{N} S_\nu / \sqrt{\sum_{\nu=1}^{N} V_\nu}$. If the families are sampled from a heterogeneous population consisting of I identified subpopulations with disease risks π_i, $i = 1,...,I$, then $\mathcal{T} = \sum_{i=1}^{I} \epsilon_i T_i / \sqrt{\sum_{i=1}^{I} \epsilon_i^2}$, where $\epsilon_i = \frac{d}{d\alpha} \log \pi_i(\alpha)|_{\alpha=\tilde{\alpha}}$ and where $\mathcal{T}_i$ is the score statistic for the subset of families from population i. In the present study, we assume that the population is homogeneous, so that $\epsilon_i \equiv \epsilon$, $i = 1,...,I$, and we may take $\epsilon = 1$ without loss of generality.

References

Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP (1997) Association mapping of disease loci, by use of a pooled DNA genomic screen. Am J Hum Genet 61:734–747

Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65:1170–1177

Cox RDR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455–464

Knapp M, Seuchter SA, Bauer MP (1993) The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. Am J Hum Genet 52:1085–1093

Little RJA, Rubin DB (1987) Statistical analysis with missing data. John Wiley & Sons, New York

Martin RB, Alda M, MacLean CJ (1998) Parental genotype reconstruction: applications of haplotype relative risk to incomplete parental data. Genet Epidemiol 15:471–490

McCullagh P, Nelder JA (1989) Generalized linear models, 2d ed. Chapman and Hall, London

Ott J (1989) Statistical properties of the haplotype relative risk. Genet Epidemiol 6:127–130

Parsian A, Todd RD, Devor EJ, O'Malley KL, Suarez BK, Reich T, Cloninger CR (1991) Alcoholism and alleles of the human D2 dopamine receptor locus: studies of association and linkage. Arch Gen Psychiatry 48:655–663

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Risch N, Teng J (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling. Genome Res 8:1273–1788

Rothman N, Caporaso NE, Wacholder S, Garcia-Closas M, Lubin JH, Marcus P, Hoover RE, et al (1999) Evaluation of interactions between environmental exposures and common genetic polymorphisms: a population-based epidemiologic perspective. Proc Am Assoc Cancer Res 40: 762–763

Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–449

Schaid DJ, Li H (1997) Genotype relative risks and association tests for nuclear families with missing parental data. Genet Epidemiol 14:1113–1118

Schaid DJ, Sommer SS (1993) Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet 53:1114–1126

Schaid DJ, Sommer SS (1994) Comparison of statistics for candidate-gene association studies using cases and parents. Am J Hum Genet 55:402–409

Self SG, Longton G, Kopecky KJ, Liang K-Y (1991) On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 47:53–62

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983–989

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Terwilliger JD, Ott J (1992) A haplotype-based "haplotype relative risk" approach to detecting allelic associations. Hum Hered 42:337–346

Thompson EA (1986) Pedigree analysis in human genetics. Johns Hopkins University Press, Baltimore

Tu I-P, Balise RR, Whittemore AS (2000) Detection of disease genes by use of family data. II. Application to nuclear families. Am J Hum Genet 66:1341–1350 (in this issue)

Whittaker JC, Lewis CM (1998) Effect of family structure on linkage tests using allelic association. Am J Hum Genet 63:889–897

Witte JS, Gauderman WJ, Thomas DC (1999) Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epidemiol 149:693–705